Clustering Examples in Multi-Dataset NLP Benchmarks with Item Response Theory

Pedro Rodriguez
me@pedro.ai

Phu Mon Htut
New York University
pmh330@nyu.edu

John P. Lalor
University of Notre Dame
john.lalor@nd.edu

Joao Sedoc
New York University
jsedoc@stern.nyu.edu

Abstract

In natural language processing, multi-dataset benchmarks for common tasks (e.g., SuperGLUE for natural language inference and MRQA for question answering) have risen in importance. Invariably, tasks and individual examples vary in difficulty. Recent analysis methods infer properties of examples such as difficulty. In particular, Item Response Theory (IRT) jointly infers example and model properties from the output of benchmark tasks (i.e., scores for each model-example pair). Therefore, it seems sensible that methods like IRT should be able to detect differences between datasets in a task. This work shows that current IRT models are not as good at identifying differences as we would expect, explain why this is difficult, and outline future directions that incorporate more (textual) signal from examples.

1 Introduction

Understanding and describing the data in natural language processing (NLP) benchmarks is crucial to ensuring their validity and reliability (Ferraro et al., 2015; Gebru et al., 2018; Bender and Friedman, 2018). This is even more important as multi-dataset task benchmarks have—for better or worse—become the norm (Raji et al., 2021). For example, SuperGLUE incorporates eight natural language inference (NLI) datasets (Wang et al., 2019), and MRQA incorporates twelve question answering (QA) datasets (Fisch et al., 2019). To better understand benchmark data, there are methods for analyzing examples in isolation (Lalor et al., 2018), characterizing a dataset's data distribution (Swayamdipta et al., 2020), using individual models to glean insight about datasets and examples (Feng et al., 2018), and using many models to do the same (Rodriguez et al., 2021; Vania et al., 2021). This paper investigates how effectively one method—Item Response Theory (IRT)—gives insight into multi-dataset benchmarks.

Outside of NLP, IRT provides insight into educational test questions (Lord et al., 1968; Baker, 2001) and political ideologies of legislators (Poole and Rosenthal, 2017). In NLP, IRT is used to identify helpful training examples (Lalor and Yu, 2020), detect errors in evaluation examples (Rodriguez et al., 2021), and estimate the future utility of examples in benchmarks (Vania et al., 2021). The goal of this paper is to identify the characteristics of multi-dataset benchmarks that IRT methods focus on. Are certain datasets easier than others? Can clustering highlight dataset or example properties?

We hypothesize that examples from similar datasets will cluster together as they should have similar IRT characteristics (such as difficulty level) compared to examples from other datasets. However, we do not see any distinct dataset-based clusters in our results. Instead, we find that IRT characteristics tend to group the examples of similar labels in the same clusters, suggesting that some label types are more difficult or more discriminating regardless of the datasets they belong to. In the rest of this paper, we describe IRT methods for benchmark analysis (§2), our clustering methods (§3), and our experimental results (§4).[1]

2 IRT for Benchmark Analysis

In this paper, we adapt IRT methods to explain *why* benchmarks examples are difficult, rather than solely assigning them difficulty values. This section describes the IRT models in our experiments and the test-bed we use in our experiments.

2.1 Item Response Theory Models

IRT is a probabilistic framework that models the likelihood that subject j (e.g., a model) answers test item i (e.g., a sentiment prediction) correctly.

[1] Code and data at www.pedro.ai/multidim-irt.

Proceedings of the Third Workshop on Insights from Negative Results in NLP, pages 100 - 112

Task	N	Datasets
Sentiment	24,620	Amazon reviews (Zhang et al., 2015), Yelp reviews,* SST-3 (Socher et al., 2013), and Dynasent Rounds 1 & 2 (Potts et al., 2021)
NLI	63,018	ANLI rounds one through three (Nie et al., 2020), HANS (McCoy et al., 2019), MNLI matched & MNLI mismatched (Williams et al., 2018), SNLI (Bowman et al., 2015), and Winogender (Rudinger et al., 2018)

*https://www.yelp.com/dataset

Table 1: Details of the datasets used in our experiments.

$$p(y_{ij} = 1 | \gamma_i, \beta_i, \lambda_i, \theta_j) = \frac{\lambda_i}{1 + e^{-\gamma_i (\theta_j - \beta_i)}} \quad (1)$$

Likelihood of correct answer for subject j on item i.

Discriminability of item i

Ability of subject j

Difficulty of item i

The likelihood of a correct response (Equation 1) is modeled as a relationship between the difficulty (β_i) of an item, its discriminability (γ_i), its feasibility (λ_i), and the subject's ability (θ_j). Typically, θ_j and β_i are unconstrained, λ_i is between zero and one, and γ_i is non-negative.

This model is a four parameter (4PL) IRT model (Equation 1) and while complex, easily simplifies to simpler models.[2] For example, when $\lambda_i = 1$ and $\gamma_i = 1$ this is a 1PL model. In this case, the difference between subject ability and item difficulty ($\theta_j - \beta_i$) determines the likelihood of a correct answer: as subject ability increases, the likelihood of a correct response increases. When only $\lambda_i = 1$, this is a 2PL model as in topic modeling experiments (§4.2). IRT parameters can also be multidimensional. In two experimental setups (§4.1 and §A), we use a 2PL model ($\lambda_i = 1$) where γ_i, β_i, and θ_j are multidimensional. We fit all models with `py-irt` (Lalor and Rodriguez, 2022).

2.2 Benchmark Data

Ideally, IRT methods should generalize across multiple datasets, tasks, and models. To accomplish this while minimizing engineering overhead, we use data from `dynabench.org` (Kiela et al., 2021)—a dynamic benchmark of multiple tasks, datasets, and model submissions (Table 1).[3] For each task, there are seven models: a majority baseline (always positive), ALBERT (Lan et al., 2020), BERT (Devlin et al., 2019), DeBERTa (He et al., 2020), FastText (Bojanowski et al., 2017; Joulin et al., 2017), RoBERTa (Liu et al., 2019), and T5 (Raffel et al., 2020). In experiments, IRT infers parameters from the subject-item (i.e., model-example) matrix where entries are one if the subject answered the item correctly and zero otherwise.

IRT analysis offers a way to assign properties like difficulty and discriminability to examples, but does little to explain why a particular example may be hard or easy. Next, we identify interpretable features that might explain IRT parameter values (e.g., label, topics, and embeddings).

3 Interpreting IRT Parameters

This section explains the methods that our experiments (§4) use to interpret IRT parameters. These methods fall into two categories: (1) methods that correlate examples' IRT parameters with dataset or label features and (2) methods that correlate derived textual information with IRT parameters (e.g., topic models or embeddings).

3.1 Multidimensional IRT Clustering

Intuitively, test instances—be they NLI examples or SAT questions—can be difficult along more than one dimension. An example might focus on testing commonsense reasoning instead of testing background knowledge. Therefore, it is sensible for IRT models to learn multidimensional parameters, but do different difficulty dimensions align with our intuitions on what might make examples easier or harder? To interpret evaluation data with multidimensional IRT, we: (1) train multidimensional IRT models,[4] (2) use t-SNE for dimensionality reduction (Poličar et al., 2019), (3) plot the resulting points in 2D space, and (4) color the points by

[2] 4PL models usually include a guessing parameter that indicates the likelihood of answering the item correctly by random guess. The guessing parameter is set to zero in our experiments.

[3] To avoid test set leakage, we use development set data.

[4] We set the dimension of the IRT model to the number of datasets per task (5 for sentiment and 8 for NLI), and the number of labels in each task (3 for both sentiment and NLI).

characteristics of each example such as the classification label or source dataset (§4.1).

3.2 Topic Models

Our next method is based on the intuition that textual information—in particular topical associations—affects example difficulty. If true, topical associations should correlate with IRT parameters. To test this, we fit a topic model to the five datasets in the Dynabench sentiment task (Table 1). To avoid having too many topics to interpret, we fit the model with five topics using the `mallet` software package (McCallum, 2002).[5] We obtain IRT parameters from a one dimensional, 2PL IRT model (Equation 1). As with multidimensional IRT, we jointly visualize an interpretable feature (topic assignment) and IRT parameter values (§4.2).

3.3 Using BERT to Predict IRT Parameters

If textual information is correlated item difficulty, then transformer models like BERT should also be able to predict IRT parameters given the item text. We test this idea by fine-tuning a BERT model (Devlin et al., 2019) with regression heads to predict the difficulty and discriminability parameters of a 4PL IRT model (Equation 1). As with the multidimensional clustering method, we also visualize embeddings from BERT-base (§4.3). The goal of our visualizations is to test: (1) how BERT embeddings change with IRT fine-tuning and (2) whether clusters correspond to interpretable instance features (e.g., label or source dataset).

4 Experiments

Next, we discuss what each interpretation method (§3) tells us about IRT parameter values.

4.1 Multidimensional IRT Clustering

Using the subject-item response matrix from Dynabench, we fit a multidimensional 2PL model, cluster with t-SNE, and color the datapoints by either dataset name or the example label.

When we run t-SNE on the difficulty parameters of a 5-dimensional 2PL model for sentiment datasets and color-code by dataset, we do not observe any distinct dataset-based clusters (Figure 1a). However, when we color-code by label, we observe more well-defined clusters, especially for the positive and negative labels (Figure 1b). This result suggests that some label types are more difficult for models to learn or more discriminating among the models regardless of which dataset they belong to. While the lack of dataset-based clustering is a negative result, label-based trends indicate consistency among items with the same label in terms of learned IRT parameters. However, the lack of breadth within a label suggests that each label can only accurately estimate a narrow range of ability levels in models.[6]

4.2 How Do Topics Relate to Item Difficulty?

We first validate that the topics inferred by the topic model (Table 2) are reasonable through manual inspection. The topic model successfully identifies at least five distinct review themes: media (e.g., movies, music), hotels, books, products, and food. Having verified that the topic model is at least reasonable, we next inspect the relationship between the highest scoring topic per example and its difficulty (Figure 3). We see that certain topics are more prevalent at different levels of difficulty; however, there is no clear delineation between topics and difficulties. This suggests that at least this topic model alone does not fully explain difficulty.[7]

4.3 How Does IRT Difficulty Influence BERT?

Figure 2 compares t-SNE visualizations of embeddings from a normal BERT model as opposed to a BERT model that is fine-tuned to predict 4PL difficulty and discriminability parameters from the sentiment task. When points are color coded by label, the embeddings of the IRT fine-tuned BERT model clearly form label-based clusters. In contrast, we do not observe clear patterns or clusters for the embeddings of the vanilla BERT model. This indicates separation of labels by IRT parameters.[8] This suggests that IRT parameters are correlated with dataset labels, and the BERT embeddings learned on IRT parameters encode label properties.

4.4 Discussion

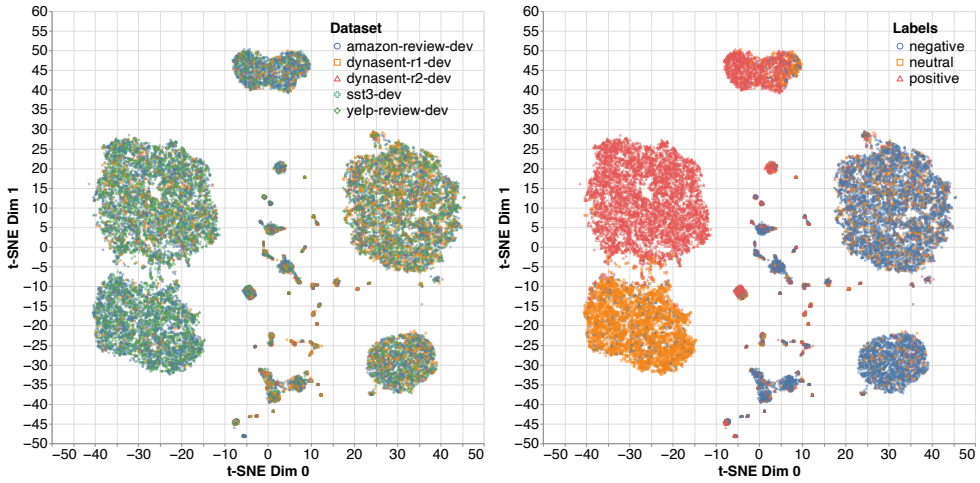
It is generally agreed that some datasets are more challenging than others. Therefore, items in the

[5] For model training, we use an optimization interval of 10 with 3,000 iterations.

[6] We performed additional clustering analyses on the sentiment and NLI datasets, varying the IRT models learned and the IRT parameters used for clustering (Appendix A). In all cases we observed more well-defined label-based clusters than dataset-based clusters.

[7] We also replicate the plot with discriminability, but do not observe any visually discernible patterns.

[8] IRT-based distributions of examples (Figure 8 in the appendices) show that there are clearer patterns with respect to IRT when we group the examples by their dataset labels.



(a) (b)

Figure 1: t-SNE visualization of sentiment datasets on the 5-dimensional 2PL IRT difficulty parameter, colored by dataset (a) and by label (b). Coloring by dataset does not result in easily discernable clusters; coloring by label produces well separated clusters for positive and neutral labels. The negation cluster is distinct but has more intruders than other labels. This suggests example label is more correlated with difficulty than source dataset.

Topic ID	Topic Words in Dynabench Sentiment Datasets
0	movie num good album music great film songs love time
1	num place time room back service people hotel didn good
2	book read story good books num reading great time characters
3	num product great good bought work time buy back price
4	num food good place great service ordered back time restaurant

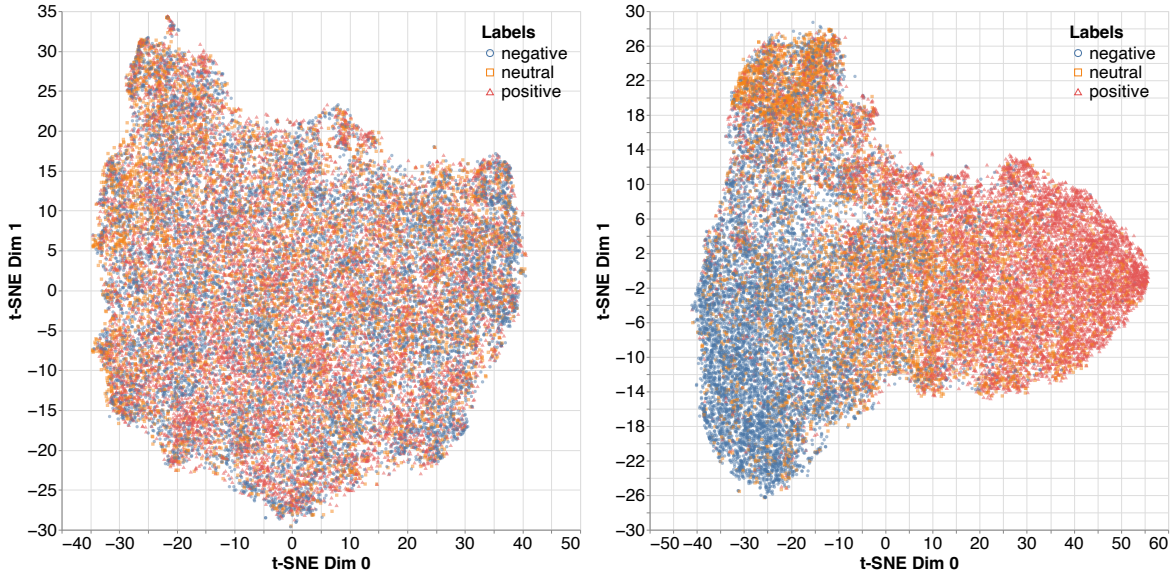
Table 2: We train a five-topic, topic model on the Dynabench sentiment data (Table 1). Topics correspond to five review themes: media, hotel, book, product, and food. Topic IDs and colors correspond to Figure 3.

same dataset should have similar IRT characteristics. However, our results indicate that benchmark datasets display more depth than breadth in terms of example IRT parameters. For a multi-dataset task such as NLI, examples clustered by IRT parameters group according to shared labels, not shared datasets. While learned latent topics show some variation across IRT difficulty, it is not clearly evident that certain topics are more difficult than others. While we cannot conclude that certain topics or datasets are more difficult than others, our results suggest that certain *labels* are.

5 Conclusion and Future Work

In this work, our expectation was that datasets would be separable by IRT-learned parameters. However, we found that clustering was more interpretable at the label level than the dataset level.

Future work in IRT should better jointly model the characteristics of NLP data as opposed to our methods that train these components in isolation. For example, it may be that the signal provided by dataset properties is second order to labels and our methods may not effectively model this (potential) multi-level relationship. Multidimensional IRT models that encode relationships between difficulty dimensions ought to better fit the data (e.g., predicting sentiment of restaurant reviews should overlap with hotel reviews, as they both involve service). If these models succeed, they should aid the interpretation of benchmarks. Lastly, as models provide more information through initiatives like Model Cards (Mitchell et al., 2019), IRT could jointly model these properties with latent ability parameters to glean insights into which differences in models yield empirical impacts.



(a) (b)

Figure 2: Clustering results for the Dynasent datasets using a BERT embeddings from a BERT model used to predict IRT parameters. 2a: Cluster by labels using untrained BERT. 2b: Cluster by labels using trained BERT. Without fine-tuning, there are no clear patterns between BERT embeddings and label. However, fine-tuning to predict IRT parameters shows clear clustering patterns between embeddings and labels. This suggests that embeddings learned to predict IRT parameters can encode the properties of dataset labels.

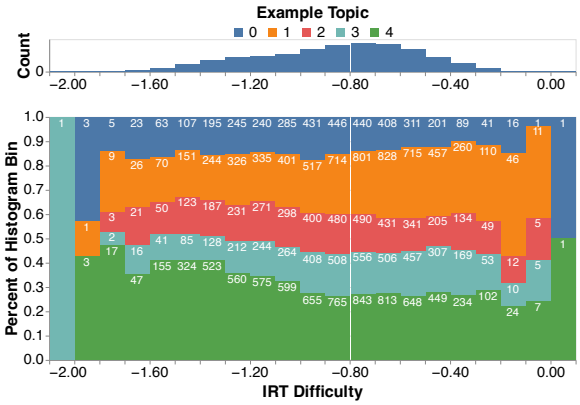


Figure 3: To observe the relationship between topics and IRT difficulty, we plot the un-normalized histogram of example difficulty (top) and the normalized difficulty partitioned by topic (bottom). Topic 4 in green (food reviews) is more prevalent with lower difficulty examples, while topic 1 in orange (hotel reviews) is more prevalent in higher difficulty examples.

References

Frank B Baker. 2001. *The Basics of Item Response Theory*. ERIC.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. 2015. A survey of current datasets for vision and language research. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019

shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2018. Datasheets for datasets.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

John P. Lalor and Pedro Rodriguez. 2022. py-irt : A scalable item response theory library for python. *arXiv preprint arXiv:2203.01282*.

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 Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

John P Lalor and Hong Yu. 2020. Dynamic data selection for curriculum learning via ability estimation. In *Findings of the Association for Computational Linguistics: EMNLP*. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

F M Lord, M R Novick, and Allan Birnbaum. 1968. Statistical theories of mental test scores.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Pavlin G. Poličar, Martin Stražar, and Blaž Zupan. 2019. openTSNE: a modular python library for t-sne dimensionality reduction and embedding.

Keith T Poole and Howard Rosenthal. 2017. *Ideology & congress: A political economic history of roll call voting*, 2 edition. Routledge, London, England.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified Text-to-Text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. In *NeurIPS: Datasets and Benchmarks Track*.

Pedro Rodriguez, Joe Barrow, Alexander 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 Association for Computational Linguistics*. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank.

In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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 Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. Comparing test sets with item response theory. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for General-Purpose language understanding systems. In *Proceedings of Advances in Neural Information Processing Systems*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage challenge corpus for sentence understanding through inference. In *Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Proceedings of Advances in Neural Information Processing Systems*.

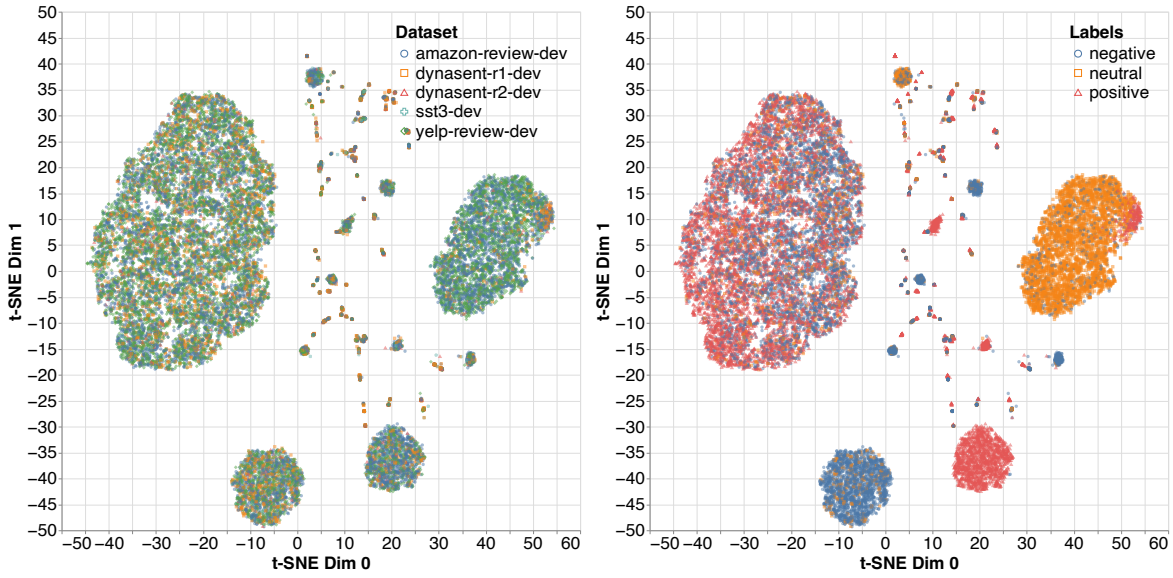
A Additional Visualizations

A.1 Dataset Based Clustering

In Figure 4a, we run t-SNE on the discriminability parameters of a 5-dimensional 2PL model learned for the Dynasent datasets and color-code by data set. We do not observe any distinct dataset-based clusters. We repeat the same visualizations using difficulty and discriminability parameters of a 3-dimensional 2PL model learned on Dynasent datasest (Figure 5a and 5c), a 3-dimensional 2PL model learned on NLI datasets (Figure 7a and 7c), and an 8-dimensional 2PL model learned on NLI datasets (Figure 6a and 6c). In all these experiments, we do not observe any distinct dataset-based cluster.

A.2 Label Based Clustering

In Figure 4b, we run t-SNE on the discriminability parameters of a 5-dimensional 2PL model learned for the Dynasent datasets and color-code by dataset labels. We repeat the same visualizations using difficulty and discriminability parameters of a 3-dimensional 2PL model learned on Dynasent datasest (Figure 5b and 5d), a 3-dimensional 2PL model learned on NLI datasets (Figure 7b and 7d), and an 8-dimensional 2PL model learned on NLI datasets (Figure 6b and 6d). In all these experiments, we observe clearer clusters compared to Section A.1.



(a) (b)

Figure 4: T-SNE visualisation of the Dynasent datasets on the discriminability parameter of a 5-dimensional 2PL model: (a) marked by dataset, (b) marked by label.

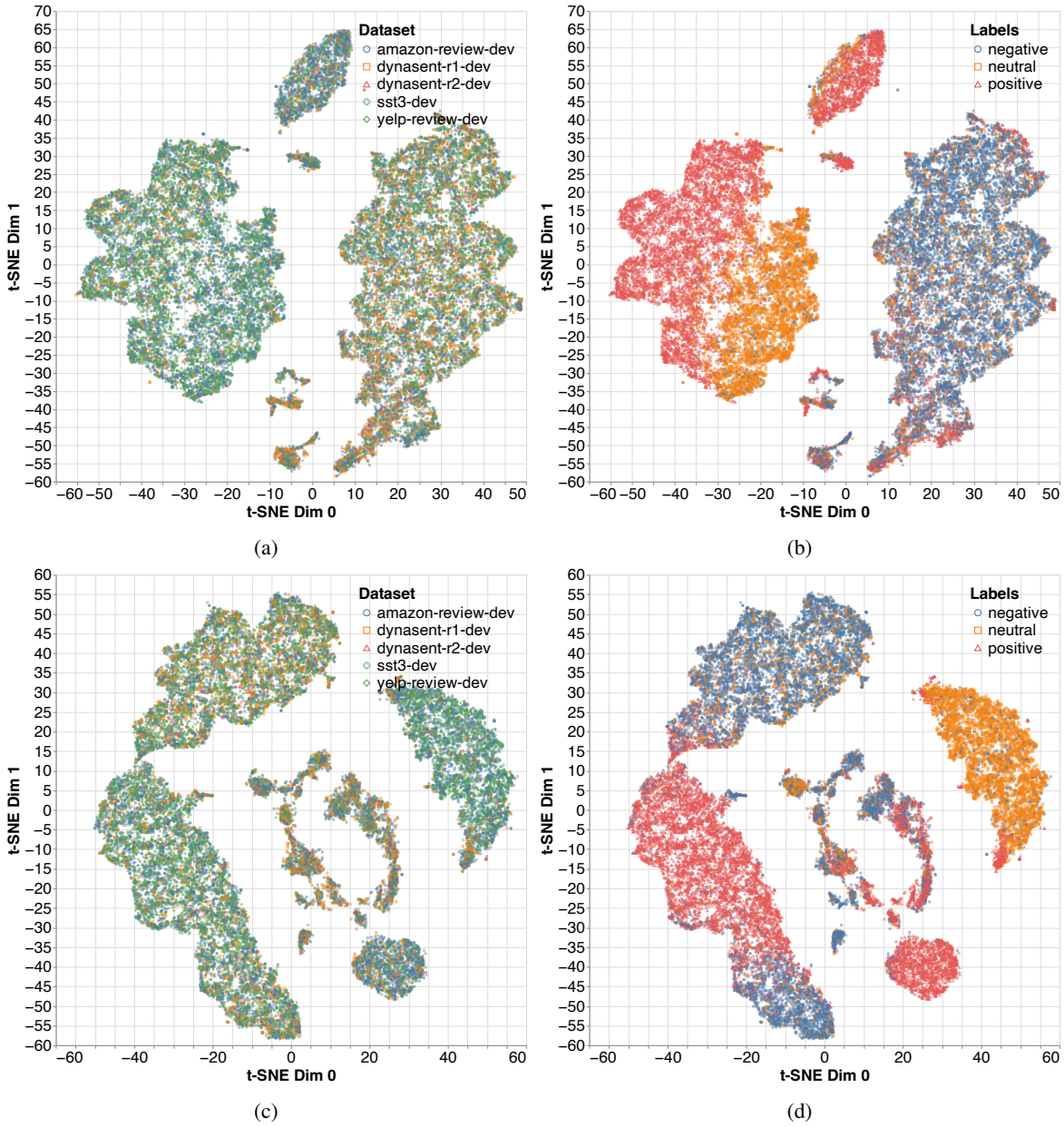
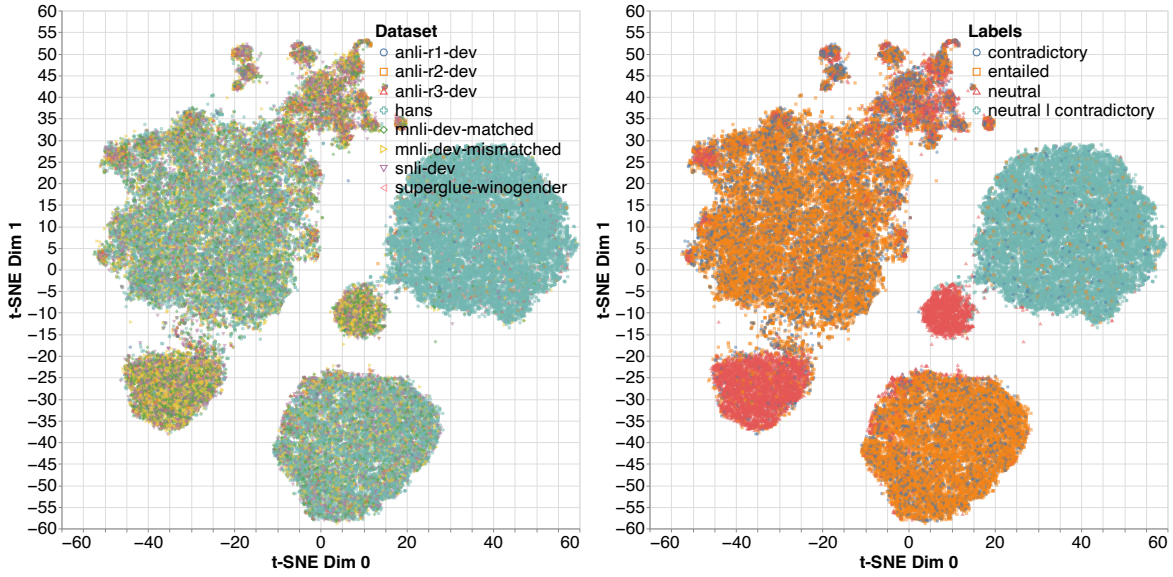
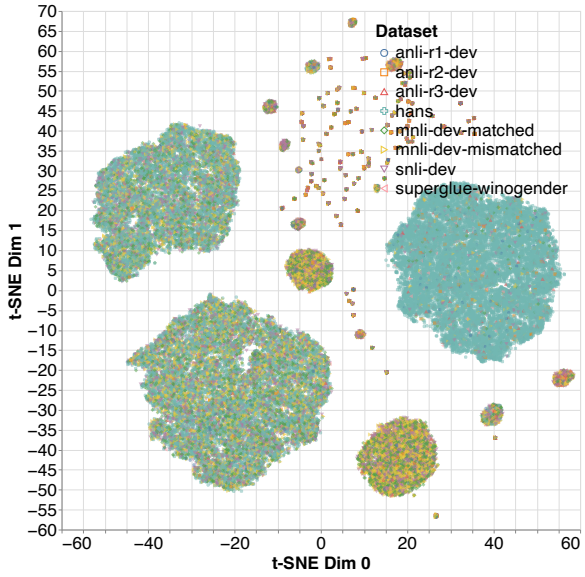


Figure 5: T-SNE visualisation of the Dynasent datasets on the parameters of a 3-dimensional 2PL model: (a) Difficulty marked by dataset, (b) Difficulty marked by label, (c) Discriminability marked by dataset, (d) Discriminability marked by label.

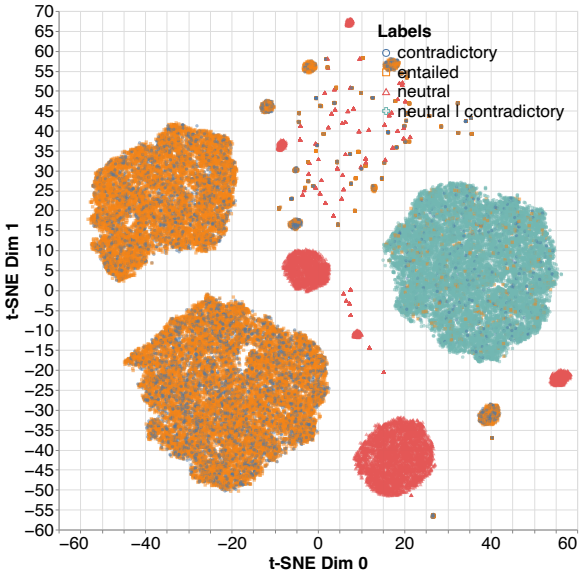


(a)

(b)

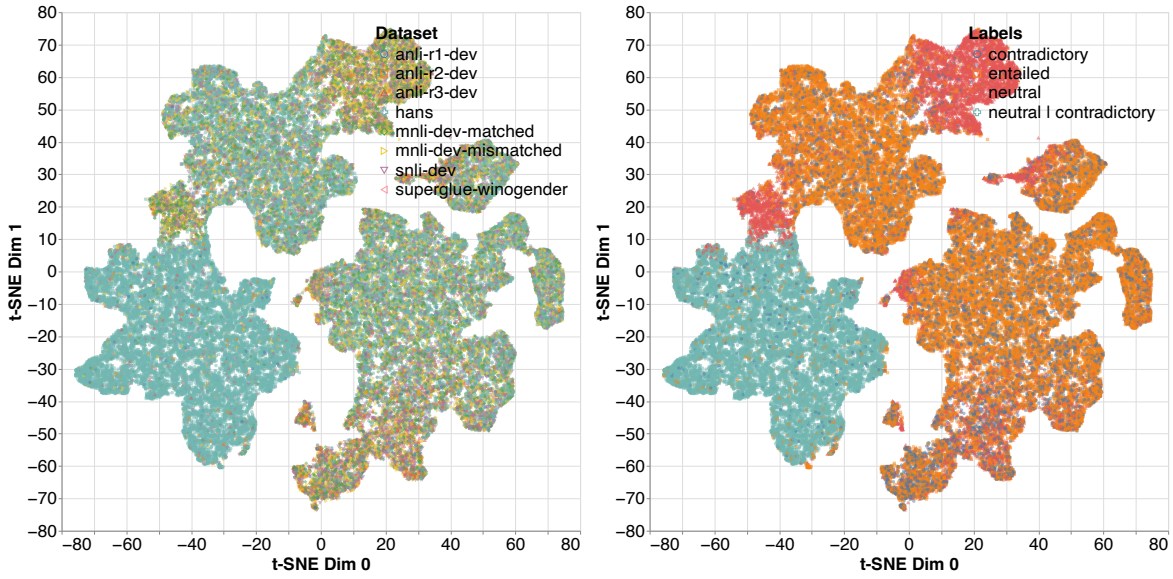


(c)



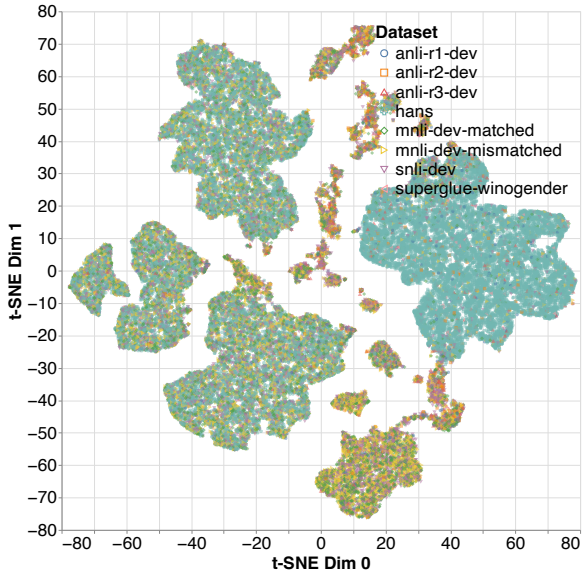
(d)

Figure 6: T-SNE visualisation of the NLI datasets on the parameters of a 8-dimensional 2PL model: (a) Difficulty marked by dataset, (b) Difficulty marked by label, (c) Discriminability marked by dataset, (d) Discriminability marked by label.

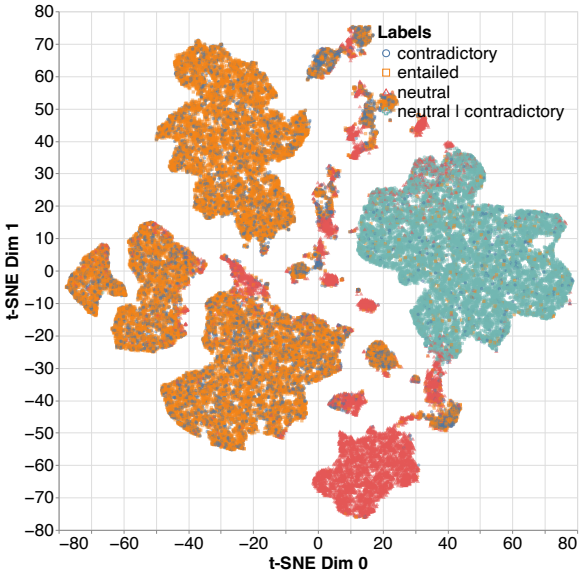


(a)

(b)

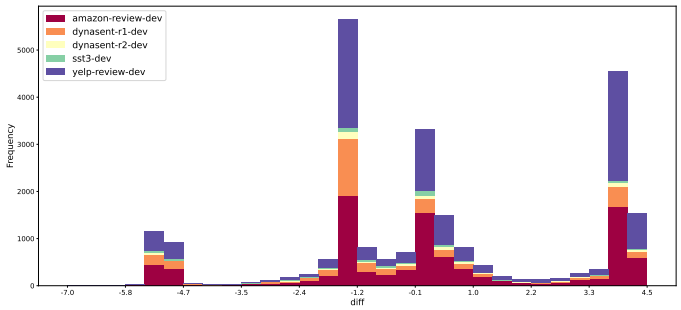


(c)

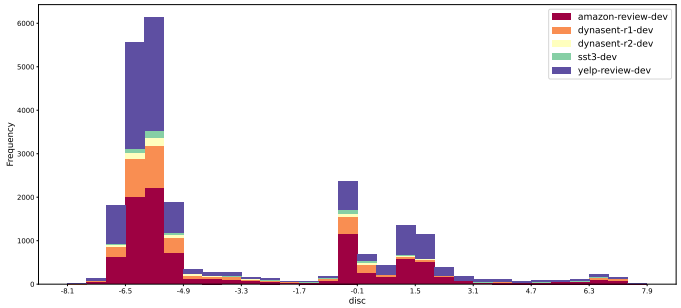


(d)

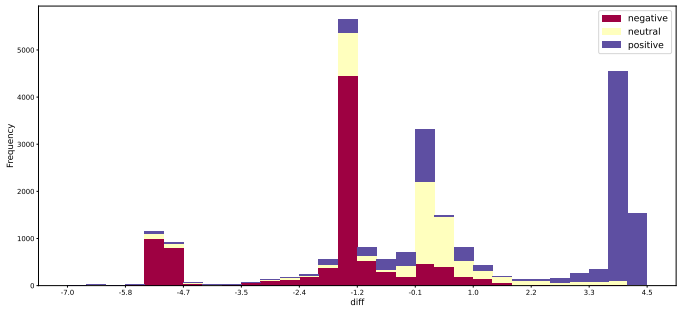
Figure 7: T-SNE visualisation of the NLI datasets on the parameters of a 3-dimensional 2PL model: (a) Difficulty marked by dataset, (b) Difficulty marked by label, (c) Discriminability marked by dataset, (d) Discriminability marked by label.



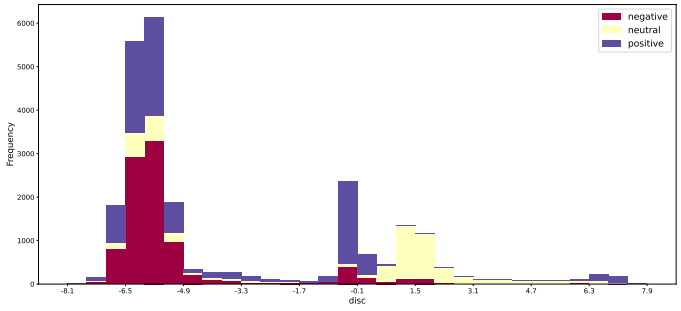
(a)



(b)



(c)



(d)

Figure 8: Distributions of examples for the sentiment datasets (3PL model): (a) Diff by dataset, (b) Disc by dataset, (c) Diff by label, (b) Disc by label.