Aiming beyond the Obvious: Identifying Non-Obvious Cases in Semantic Similarity Datasets

Nicole Peinelt[1,2] and **Maria Liakata**[1,2] and **Dong Nguyen**[1,3]
[1]The Alan Turing Institute, London, UK
[2]University of Warwick, Coventry, UK
[3]Utrecht University, Utrecht, The Netherlands
{n.peinelt, m.liakata}@warwick.ac.uk, dnguyen@turing.ac.uk

Abstract

Existing datasets for scoring text pairs in terms of semantic similarity contain instances whose resolution differs according to the degree of difficulty. This paper proposes to distinguish obvious from non-obvious text pairs based on superficial lexical overlap and ground-truth labels. We characterise existing datasets in terms of containing difficult cases and find that recently proposed models struggle to capture the non-obvious cases of semantic similarity. We describe metrics that emphasise cases of similarity which require more complex inference and propose that these are used for evaluating systems for semantic similarity.

id	case	documents
160174	P_o	what's the origin of the word o'clock? what is the origin of the word o'clock?
115695	P_n	which is the best way to learn coding? how do you learn to program?
193190	N_o	what are the range of careers in biotechnology in indonesia? how do you tenderize beef stew meat?
268368	N_n	what is meant by 'e' in mathematics? what is meant by mathematics?

Table 1: Examples for difficulty cases from the development set of the Quora dataset. o=obvious, n=non-obvious, N=negative label, P=positive label

1 Introduction

Modelling semantic similarity between a pair of texts is a fundamental task in NLP with a wide range of applications (Baudiš et al., 2016). One area of active research is Community Question Answering (CQA) (Nakov et al., 2017; Bonadiman et al., 2017), which is concerned with the automatic answering of questions based on user generated content from Q&A websites (e.g. StackExchange) and requires modelling the semantic similarity between question and answer pairs. Another well-studied task is paraphrase detection (Socher et al., 2011; He et al., 2015; Tomar et al., 2017), which models the semantic equivalence between a pair of sentences.

Evaluation for such tasks has primarily focused on metrics, such as mean average precision (MAP), F1 or accuracy, which give equal weights to all examples, regardless of their difficulty. However, as illustrated by the examples in Table 1, not all items within text pair similarity datasets are equally difficult to resolve.

Recent work has shown the need to better understand limitations of current models and datasets in natural language understanding (Wadhwa et al., 2018a; Rajpurkar et al., 2018). For example, Kaushik and Lipton (2018) showed that models sometimes exploit dataset properties to achieve high performance even when crucial task information is withheld, and Gururangan et al. (2018) demonstrated that model performance is inflated by annotation artefacts in natural language inference tasks.

In this paper, we analyse current datasets and recently proposed models by focusing on item difficulty based on shallow lexical overlap. Rodrigues et al. (2018) found declarative CQA sentence pairs to be more difficult to resolve than interrogative pairs as the latter contain more cases of superficial overlap. In addition, Wadhwa et al. (2018b) showed that competitive neural reading comprehension models are susceptible to shallow patterns (e.g. lexical overlap). Our study digs deeper into these findings to investigate the properties of current text pair similarity datasets with respect to different levels of difficulty and evaluates models based on how well they can resolve difficult cases.

We make the following contributions:

1. We propose a criterion to distinguish between obvious and non-obvious examples in text

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2792–2798
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

pair similarity datasets (section 4).

2. We characterise current datasets in terms of the extent to which they contain obvious vs. non-obvious items (section 4).

3. We propose alternative evaluation metrics based on example difficulty (section 5) and provide a reference implementation at https://github.com/wuningxi/LexSim.

2 Datasets and Tasks

We selected well-known benchmark datasets differing in size (small vs. large), document length (single sentence vs. multi-sentence), document types (declarative vs. interrogative) and tasks (answer ranking vs. paraphrase detection vs. similarity scoring), see Table 2.

SemEval The SemEval Community Question Answering (CQA) dataset (Nakov et al., 2015, 2016, 2017) contains posts from the online forum Qatar Living. The task is to rank relevant posts above non-relevant ones. Each subtask involves an initial post and 10 possibly relevant posts with binary annotations. Task A contains questions and comments from the same thread, task B involves question paraphrases, and task C is similar to A but contains comments from an external thread.

MSRP The Microsoft Research Paraphrase corpus (MSRP) is a popular paraphrase detection dataset, consisting of pairs of sentences with binary judgments (Dolan and Brockett, 2005).

Name	Task	Type	Size
SemEval	(A) answer ranking	rank	26K
	(B) paraphrase ranking	rank	4K
	(C) answer ranking	rank	47K
Quora	paraphrase detection	class	404K
MSRP	paraphrase detection	class	5K
STS	similarity scoring	regr	8K

Table 2: Selected text pair similarity data sets. Size as number of text pairs. rank=ranking task, class=classification task, regr=regression task.

Quora The Quora duplicate questions dataset contains a large number of question pairs with binary labels[1]. The task is to predict whether two questions are paraphrases, similar to Task B of SemEval, but it is framed as a classification rather than a ranking problem. We use the same training / development / test set partition as Wang et al. (2017).

STS The Semantic Textual Similarity Benchmark (STS) dataset (Cer et al., 2017) consists of a selection of STS SemEval shared tasks (2012-2017). It contains sentence pairs annotated with continuous semantic relatedness scores on a scale from 0 (low similarity) to 5 (high similarity).

In this paper, we focus on predicting the semantic similarity between two text snippets in a binary classification scenario, as the ranking scenario is only applicable to some of the datasets. Binary labels are already provided for all tasks except for

[1]https://engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning

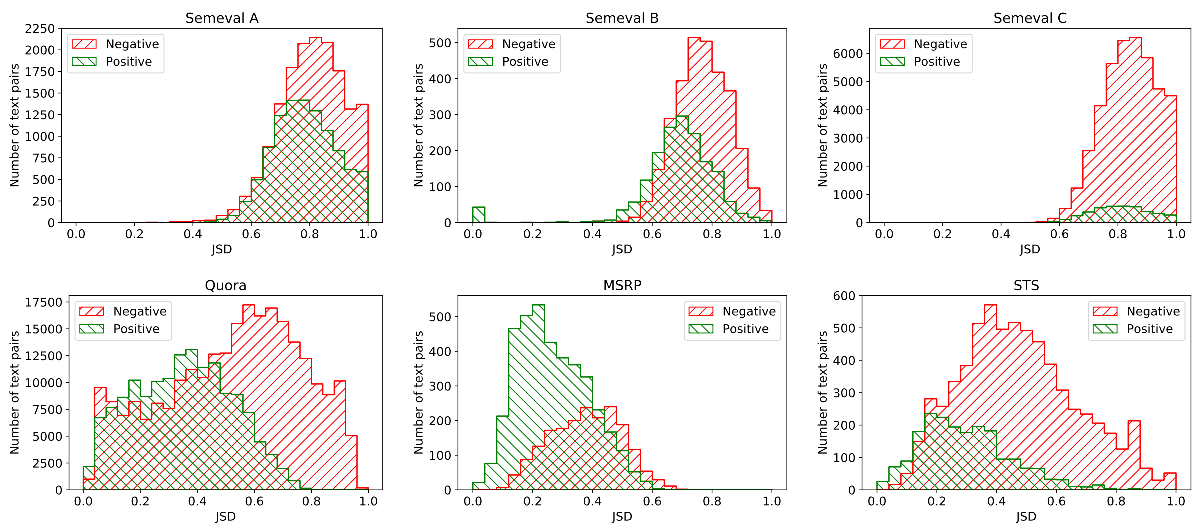

Figure 1: Lexical divergence distribution by labels across datasets. JSD=Jensen-Shannon divergence.

STS. In the case of STS, we convert the scores into binary labels. Based on the description of the relatedness scores in Cer et al. (2017), we assign a positive label if relatedness ≥ 4 and a negative one otherwise to use a similar criterion as in the other datasets.

3 Lexical divergence in current datasets

To characterise the datasets, we represent the text pairs as two distributions over words and measure their lexical divergence using Jensen-Shannon divergence (JSD) (Lin, 1991).[2] Figure 1 shows the entire JSD distribution by label for each dataset.

The datasets differ with respect to the degree of lexical divergence they contain: The three SemEval CQA datasets show a high degree of lexical divergence (majority > 0.5), especially in the external QA scenario (task C). Text pairs in MSRP tend to have low-medium JSD scores (majority < 0.6), while items in Quora and STS show the widest range of lexical divergence (see also Appendix A). Overall, pairs with negative labels tend to have higher JSD scores than pairs with positive labels. Especially in Quora, MSRP and STS, distinct distributions emerge for positive vs. negative labels, providing direct clues for label assignment.

4 Distinguishing between obvious and non-obvious examples

As shown, pairs with high lexical divergence tend to have a negative label in the above datasets (e.g. N_o in Table 1), while low lexical divergence is associated with a positive label (e.g. P_o in Table 1). Intuitively, these are cases which should be relatively easy to identify. More difficult are text pairs with a positive label but high lexical divergence (e.g. P_n in Table 1), or a negative label despite low lexical divergence (e.g. N_n in Table 1). We use Table 3 to categorise cases in terms of their difficulty level.

	positive label	negative label
low div	obvious pos (P_o)	non-obvious neg (N_n)
high div	non-obvious pos (P_n)	obvious neg (N_o)

Table 3: Defining obvious and non-obvious similarity cases based on labels and lexical overlap.

[2] We also calculated set-based similarity metrics (Jaccard Index and Dice Coefficient) and found consistent results with JSD, but give preference to the distribution-based metric which is more natural for text. Due to space restrictions, we only report JSD in this paper.

	Fleiss' Kappa	Avg. time per pair	Instances
P_o	0.6429	11.58s	35
P_n	0.0878	11.68s	15
N_o	0.3886	12.50s	34
N_n	0.0892	13.83s	16
total	0.6267	12.27s	100

Table 4: Statistics for manual annotation on Quora. o=obvious, n=non-obvious, N=negative, P=positive

	SemEval			Quora	MSRP	STS
	A	B	C			
P_o	5893	1162	2492	107612	2398	1597
P_n	4428	531	1590	41691	1502	409
N_o	8842	1843	22155	160410	1398	3900
N_n	7377	1213	21253	94632	503	2719
o	56	63	52	66	65	64
m	0.80	0.79	0.82	0.53	0.52	0.52

Table 5: Difficulty case splits across datasets (train, dev and test combined). o=obvious, m=median JSD.

Pairs are categorised into high and low lexical divergence categories by comparing their JSD score to the median of the entire JSD distribution in order to account for differences between datasets (>median: high div, ≤median: low div). To verify if this automatic difficulty distinction corresponds with real-world difficulty, the authors of the study annotated the semantic relatedness of 100 random pairs from the Quora development set and measured inter-annotator agreement based on Fleiss' Kappa. The agreement for non-obvious cases (P_n and N_n) is significantly lower (p-value$<$ 0.01 with permutation test) than for obvious cases (P_o and N_o) and the average annotation time per item is longer for non-obvious cases (Table 4), confirming the validity of this distinction.

Table 5 shows the number of instances in the four cases across datasets. In all of the analysed datasets, there are more obvious positives (P_o) than non-obvious positives (P_n) and more obvious negatives (N_o) than non-obvious negatives (N_n). All obvious cases combined (P_o+N_o) make up more than 50% of pairs across all datasets.

5 Evaluating model predictions based on difficulty

We now use this categorisation for the purpose of model evaluation (Tables 6-8).[3] We calculate the

[3] Due to the lack of openly available model prediction files, we only present our analysis for the Se-

	KeLP	Beihang MSRA	IIT UHH	ECNU	bunji	EICA	Swiss Alps	FuRong Wang	FA3L	Snow Man	ran-dom
TPR_o	0.652	**1.000**	0.800	0.790	0.681	0.328	0.333	0.562	0.691	0.677	0.501
TPR_n	0.496	**1.000**	0.676	0.636	0.575	0.269	0.223	0.399	0.478	0.469	0.499
TNR_o	0.909	0.000	0.731	0.877	0.894	0.959	**0.984**	0.913	0.787	0.900	0.515
TNR_n	0.908	0.000	0.676	0.820	0.851	**0.953**	0.950	0.892	0.751	0.757	0.536
$F1_o$	0.751	0.682	0.781	**0.829**	0.765	0.480	0.494	0.684	0.731	0.765	0.513
$F1_n$	0.628	0.686	0.686	**0.707**	0.672	0.410	0.352	0.533	0.560	0.555	0.519
F1	0.698	0.684	0.739	**0.777**	0.725	0.450	0.433	0.621	0.659	0.673	0.516
MAP	**0.884**	0.882	0.869	0.867	0.866	0.865	0.862	0.843	0.834	0.818	0.623

Table 6: Proposed evaluation metrics for top 10 primary submissions on SemEval Task A. The systems are ordered in columns according to their MAP ranking. Bold indicates the highest value for each metric. We indicate the 2^{nd} and 3^{rd} systems based on $F1_n$ and F1.

	Sim Bow	LearningTo Question	KeLP	Talla	Beihang MSRA	NLM NIH	Uin-suska TiTech	IIT UHH	SCIR QA	FA3L	ran-dom
TPR_o	0.976	**1.000**	0.920	0.760	**1.000**	0.880	0.752	0.704	0.912	0.448	0.552
TPR_n	0.842	**1.000**	0.632	0.763	**1.000**	0.500	0.421	0.737	0.842	0.263	0.395
TNR_o	0.609	0.000	0.831	0.684	0.000	0.841	0.858	0.682	0.709	**0.861**	0.495
TNR_n	0.197	0.000	0.432	0.467	0.000	0.397	0.552	0.403	0.352	**0.756**	0.521
$F1_o$	0.604	0.383	**0.746**	0.548	0.383	0.736	0.681	0.516	0.641	0.473	0.348
$F1_n$	0.198	0.195	0.199	**0.247**	0.195	0.154	0.164	0.221	0.234	0.160	0.147
F1	0.424	0.312	**0.506**	0.426	0.312	0.473	0.467	0.390	0.464	0.365	0.280
MAP	**0.472**	0.469	0.467	0.457	0.448	0.446	0.434	0.431	0.427	0.422	0.298

Table 7: Proposed evaluation metrics for top 10 primary submissions on SemEval Task B.

true positive rate TPR (for P_o and P_n) and true negative rate TNR (for N_o and N_n) to analyse model performance within each difficulty category. In the three SemEval 2017 CQA tasks, all systems perform worse on the hard cases compared to the obvious cases ($TPR_n < TPR_o$ and $TNR_n < TNR_o$), while there are only minor changes in the random baseline which predicts all classes with equal probability. To compare how well models do on obvious vs. non-obvious cases overall, we compute F1 scores for obvious cases (P_o and N_o) as $F1_o$ and non-obvious cases (P_n and N_n) as $F1_n$ separately. This is necessary as the high percentage of obvious cases (observed in section 4) can inflate the overall F1 score. $F1_n$ scores are consistently lower than the $F1_o$ scores. This difference is especially pronounced in Task B, which contained the highest proportion of obvious cases (62%) of the SemEval tasks. Using the non-obvious F1 scores results in a different ranking compared to the official SemEval evaluation metrics (F1 or MAP), even resulting in a change in the highest ranked system in Task B (Talla instead of KeLP or SimBow) and C (KeLP instead of bunji or IIT-UHH).

	IIT UHH	bunji	KeLP	EICA	ran-dom
TPR_o	0.570	0.246	**0.911**	0.006	0.520
TPR_n	0.358	0.045	**0.836**	0.000	0.433
TNR_o	0.898	0.991	0.720	**0.998**	0.502
TNR_n	0.779	0.965	0.538	**0.999**	0.502
$F1_o$	0.283	**0.339**	0.209	0.011	0.076
$F1_n$	0.047	0.028	**0.054**	0.000	0.027
F1	0.144	**0.197**	0.121	0.008	0.053
MAP	**0.155**	0.147	0.144	0.135	0.058

Table 8: Proposed evaluation metrics for top 4 primary submissions on SemEval Task C.

6 Conclusion

We present an automated criterion for automatically distinguishing between easy and difficult items in text pair similarity prediction tasks. We find that more than 50% of cases in current datasets are relatively obvious. Recently proposed models perform significantly worse on non-obvious cases compared to obvious cases. In or-

mEval CQA Tasks based on prediction files obtained from http://alt.qcri.org/semeval2017/task3/index.php?id=results.

der to encourage the development of models that perform well on difficult items, we propose to use non-obvious F1 scores ($F1_n$) as a complementary ranking metric for model evaluation. We also recommend publishing prediction files along with models to facilitate error analysis.

Acknowledgments

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Petr Baudiš, Jan Pichl, Tomáš Vyskočil, and Jan Šedivý. 2016. Sentence Pair Scoring: Towards Unified Framework for Text Comprehension. *arXiv preprint arXiv:1603.06127*.

Daniele Bonadiman, Antonio Uva, and Alessandro Moschitti. 2017. Effective Shared Representations with Multitask Learning for Community Question Answering. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 726–732, Valencia, Spain. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP@IJCNLP)*, pages 9–16, Jeju Island, Korea. Asian Federation of Natural Language Processing.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Hua He, Kevin Gimpel, and Jimmy Lin. 2015. Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1576–1586, Lisbon, Portugal. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C Lipton. 2018. How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Jianhua Lin. 1991. Divergence Measures based on the Shannon Entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Preslav Nakov, Doris Hoogeveen, Llúis Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 Task 3: Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval@ACL 2017)*, pages 27–48, Vancouver, Canada. Association for Computational Linguistics.

Preslav Nakov, Lluis Marquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, James Glass, and Bilal Randeree. 2016. SemEval-2016 Task 3: Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT 2016)*, pages 525–545, San Diego, California. Association for Computational Linguistics.

Preslav Nakov, Lluis Marquez, Magdy Walid, Alessandro Moschitti, James Glass, and Bilal Randeree. 2015. SemEval-2015 task 3: Answer Selection in Community Question Answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT 2015)*, pages 269–281, Denver, Colorado. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Joao Rodrigues, Chakaveh Saedi, Antonio Branco, and Joao Silva. 2018. Semantic Equivalence Detection: Are Interrogatives Harder than Declaratives? In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 3248–3253, Miyazaki, Japan. European Language Resources Association.

Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. 2011. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 801–809, Granada, Spain.

Gaurav Singh Tomar, Thyago Duque, Oscar Täckström, Jakob Uszkoreit, and Dipanjan Das. 2017. Neural Paraphrase Identification of Questions with Noisy Pretraining. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 142–147, Copenhagen, Denmark. Association for Computational Linguistics.

Soumya Wadhwa, Khyathi Raghavi Chandu, and Eric Nyberg. 2018a. Comparative Analysis of Neural QA models on SQuAD. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 89–97, Melbourne, Australia. Association for Computational Linguistics.

Soumya Wadhwa, Varsha Embar, Matthias Grabmair, and Eric Nyberg. 2018b. Towards Inference-Oriented Reading Comprehension: ParallelQA. In *Proceedings of the Workshop on Generalization in the Age of Deep Learning*, pages 1–7, New Orleans, Louisiana. Association for Computational Linguistics.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral Multi-Perspective Matching for Natural Language Sentences. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4144–4150, Melbourne, Australia.

A Appendix

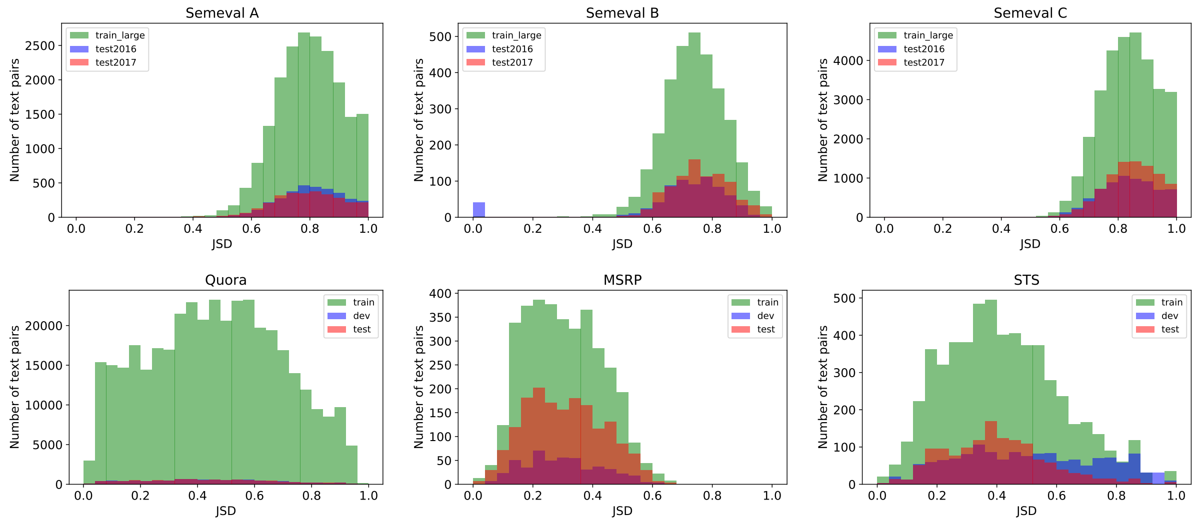

Figure 2: Lexical divergence distribution by training, development and test set across different semantic similarity datasets. JSD=Jensen-Shannon divergence.