INVESTIGATING VARIABILITY IN DATASETS AND IT'S IMPACT ON MODEL ROBUSTNESS

Abhishek Anand, Muyuan He
{anandabh,muyuanhe}@usc.edu

ABSTRACT

In the realm of machine learning, the integrity of training datasets is paramount for the development of robust and generalizable models. This paper addresses critical challenges in dataset variability and model generalization across Natural Language Processing (NLP) tasks. Our primary objective is twofold: Firstly, we conduct a comprehensive assessment of datasets in NLP, focusing on measuring the variability present in them. We propose a novel approach through explained variance curves using PCA to evaluate the variety present in popular NLP datasets for 2 tasks - Sentiment Classification and Natural Language Inference. Secondly, we empirically demonstrate the tendency of models to overfit on low-variability datasets, which leads to poor generalization in real-world scenarios. We investigate the hypothesis that semantic duplicates within these datasets do not substantially contribute to model performance and propose a methodology aimed at training models with enhanced generalization abilities. Through our research, we seek to elucidate the impact of dataset variability on model performance and advance the field's understanding of effective training methodologies. Our findings aim to contribute to the development of machine learning models that are not only high-performing with reduced training time but also robust and adaptable to a variety of real-world applications. The repository of this project is available at Github.

1 INTRODUCTION

In machine learning, a model's success is deeply connected to the variability of its training dataset. Each data point is crucial, shaping the model's abilities. However, challenges like noisy or mislabeled samples, and semantic duplicates that offer limited new information, can impair performance. These issues are especially significant in benchmark testing, often used to evaluate model generalization. Such benchmarks are evaluated against in-distribution test sets and fail to check model's generalization capabilities. Our study addresses these concerns by analyzing benchmark datasets for NLP tasks, examining the factor of variability in datasets and its effect on model performance. We hypothesize that models might excel on test subsets of the same dataset but struggle with different datasets for similar tasks. To counter this, we employ PCA-based analysis technique to identify appropriate subset size for a sampling technique to achieve best in-distribution and out-of-distribution performance, aiming to improve learning and generalization. We aim to deepen the understanding of dataset composition and guide the development of more robust, generalizable machine learning models.

2 RELATED WORK

Much of the existing work in the language modelling domain on variability in datasets has been on selecting representative subsets of data that can achieve performance close to training on the full dataset. Multiple methods have been explored by previous works on selecting such a subset. For example, Zhang et al. (2016) proposed a method RepExtract that utilized K-Means on tf-idf vectors to extract a representative subset. Abbas et al. (2023) utilized clustering of pre-trained embeddings to remove semantic duplicates present in the dataset, achieving full set accuracy and better generalization performance. Suzuki et al. (2023) used a likelihood difference score to construct a RepSet utilizing a general and in-domain model. Other works have explored removing exact duplicates from language datasets. In Raffel et al. (2023), C4 text corpus used for training the T5 model was deduplicated by removing repeating occurrences of three-sentence spans. Rae et al. (2021) applied MinHash based deduplication for Gopher model and showed that training on the dataset showed improved performance on the validation set. Exact match deduplication cleans datasets but lacks semantic duplicate handling, while clustering methods, effective for semantics, require training to determine subset size and model comparison with the full dataset. Finally, Jain et al. (2023) presents a technique in which they train a model SubSelNet for subset selection.

On the other hand, there have been techniques for computing value a particular sample provides to training and thus, building a subset of the most informative samples. For example, Ethayarajh et al. (2022) introduces a method to calculate the usable information of subsets and individual samples to build informative subsets. Killamsetty et al. (2021) presents a method to extract a subset whose gradient matches that of the training set. While both the methods demonstrate promising results on in-distribution and out-of-distribution performance, they required training on the full dataset to get the required metrics.

Finally, there has been previous work on testing robustness of models. For example, Talman & Chatzikyriakidis (2018) performed generalization testing of NLI models by evaluating them across benchmarks to demonstrate lack of robustness. Singh et al. (2021) replaces words with semantically similar meanings to test model robustness.

3 PROBLEM FORMULATION

Our research is structured around a series of meticulously defined objectives, each contributing to a comprehensive understanding of dataset variability and model generalization in the context of Natural Language Processing (NLP) tasks.

3.1 DATASET VARIABILITY ASSESSMENT

Our first objective involves an assessment of the variability in datasets used in various NLP tasks. This includes identifying and quantifying the extent of redundant samples within these datasets. We aim to show empirically that a measure of variability in the datasets which evaluates the extent of semantically redundant samples present in them. For that purpose, we develop a measure of variety in a dataset that doesn't require training by randomly sampling subsets of different sizes and measure the variance in the entire set that can be explained by the subset.

3.2 MODEL GENERALIZATION AND OVERFITTING

A key aspect of our research is to empirically demonstrate the propensity of models to overfit to low-variability datasets, leading to suboptimal generalization in real-world applications. We will investigate this hypothesis that the presence of semantic duplicates in datasets does not significantly improve model performance beyond what could be achieved without them. Moreover, we demonstrate the phenomena of overfitting to a data distribution by training on the complete dataset with low variability which results in the model learning spurious relations and hence, resulting in degraded generalization performance. The research will also propose and evaluate a methodology focused on training models that are better equipped to generalize across various tasks, thereby overcoming the limitations of overfitting.

Through these objectives, our research endeavors to not only illuminate the intricacies of dataset composition and its impact on model performance but also to advance methodologies that help construct better datasets which enhance the robustness and generalizability of machine learning models across diverse applications.

Sentiment Classification

Dataset Name	**Training Size**	**Testing Size**	**Validation Size**	**Total Size**
IMDB	25000	25000	—	50000
SST2	67349	1821	872	70042
Tweet_eval	45615	12284	2000	59899

Natural Language Inference

Dataset Name	**Training Size**	**Testing Size**	**Validation Size**	**Total Size**
ANLI	100459	1200	1200	102859
MultiNLI	392702	9832	9815	412349
SNLI	550,152	10,000	10,000	570152

Table 1: Training datasets for measuring variance for Sentiment Classification and Natural Language Inference. Datasets: IMDB Maas et al. (2011), SST2 Socher et al. (2013), Tweet _eval Rosenthal et al. (2017), ANLI Nie et al. (2020), MultiNLI Williams et al. (2018), SNLI Bowman, Samuel R. and Angeli, Gabor and Potts, Christopher, and Manning, Christopher D. (2015)

4 METHODOLOGY

4.1 DATASET VARIABILITY

In this section, we outline the method to determine the variability present in a dataset by plotting the explained variance curve. The steps are defined as follows –

1. Consider a dataset D with n samples and m features. We randomly sample 10 subsets from the dataset iteratively with incrementing sizes from 10% to 100% with an interval of 10%..
2. For each subset i, we construct a PCA model p_i with top k components where k ¡ m.
3. Using each subset's p_i, we transform D to the components space to get D^i and then perform inverse transform of D^i to attempt to reconstruct the original data in m feature space. We denote this representation as D^i_m.
4. For each subset, we compute the R-squared (R2) score r_i between D^i_m and D, which gives the percentage of variance in the complete dataset D, that can be explained by D^i_m.
5. We visualize the explained variance curve by plotting each subset size against its r_i value.

We utilize Principal Component Analysis (PCA) as the fundamental unit of our variability measure as its components represent the axes in the dataset with maximum variability sorted in decreasing order.

4.2 MODELLING

We train a model for each subset from the dataset variability analysis phase for the task T represented by the dataset D to evaluate the following key factors –

1. Evaluate accuracy improvement by training on the entire dataset D versus its subsets, checking in-distribution performance on the test set.
2. Assess out-of-distribution performance for each subset model to gauge robustness and detect potential overfitting, considering the model's performance on other datasets of the same task and calculating the mean.
3. Determine correlation between subset models' generalization performance and its respective explained variance curve.

4.3 EXPERIMENTAL SETUP

We consider 2 Natural Language Processing tasks – Sentiment Classification and Natural Language Inference for our experiments. For each task we perform our analysis on 3 datasets described in Table 1 to present a fair overview of the results. We fine-tune Devlin et al. (2018) bert-base-uncased for our modelling phase. To construct features from sentences for variability analysis, we

Sentiment Classification

Dataset Name	Testing Size
Rotten Tomatoes	1066
Yelp Reviews	38000
Amazon Kindle Reviews	12000
Twitter	53879
Reddit	24107
Finance Sentiment	2712

Natural Language Inference

Dataset Name	Testing Size
ANLI R1	1000
ANLI R2	1000
SICK	4906
SemEval 2014 Task 1	4927

Table 2: Out of distribution datasets for model robustness evaluation for Sentiment Classification and Natural Language Inference. Datasets: Rotten Tomatoes Pang & Lee (2005), Yelp Reviews Zhang et al. (2015), Amazon Kindle Reviews He & McAuley (2016), Twitter Gowda et al. (2019), Reddit Gowda et al. (2019), Finance Sentiment Malo et al. (2013), ANLI R1 Nie et al. (2020), ANLI R2 Nie et al. (2020), SICK Marelli et al. (2014b), SemEval 2014 Task 1 Marelli et al. (2014a)

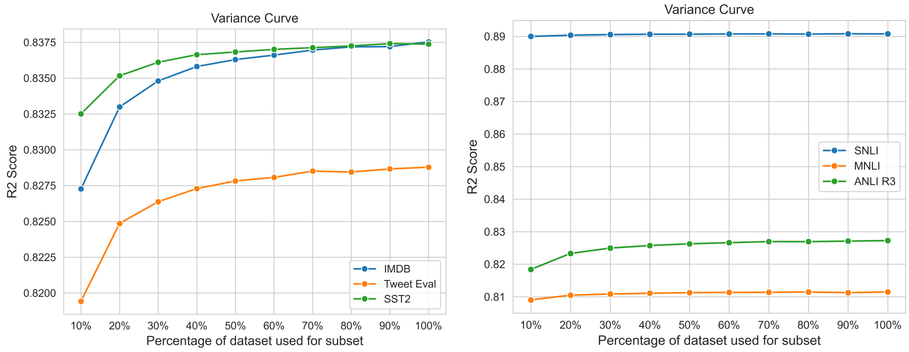

Figure 1: Explained Variance Curves for the datasets. First plot shows the results for Sentiment Classification and the second plot shows the results for Natural Language Inference datasets. On the x-axis we have the percentage of dataset used and on the y-axis we have the R2 score.

use the Reimers & Gurevych (2019) paraphrase-MiniLM-L6-v2 pre-trained model from sentence-transformers package to compute sentence embeddings of 384 dimension. For sentiment classification, direct text was used as input and for natural language inference, concatenation of premise and hypothesis text was used as input. We utilize sklearn library for performing PCA, where we use the top 100 components.

Subsequently, models were trained on all the subsets of the datasets included in the analysis. In total, we train 60 models, one each for a subset of a dataset. To measure the in-distribution performance we evaluate the models on their test sets. To measure the out-of-distribution performance for evaluating model robustness, we test the models on the other 2 datasets of the same task. Additionally, to get a representative score of model robustness, for sentiment classification we employ 6 more datasets and for natural language inference we employ 4 more datasets for out-of-distribution evaluation. These datasets are listed in Table 2.

5 RESULTS AND DISCUSSION

The analysis of explained variance curves in Figure 1 and Figure 1 reveals a consistent trend showing that the proportion of variability in the complete dataset explained by the subsets initially rises as we increase the subset size, eventually reaching a plateau regardless of the increasing sizes, indicating

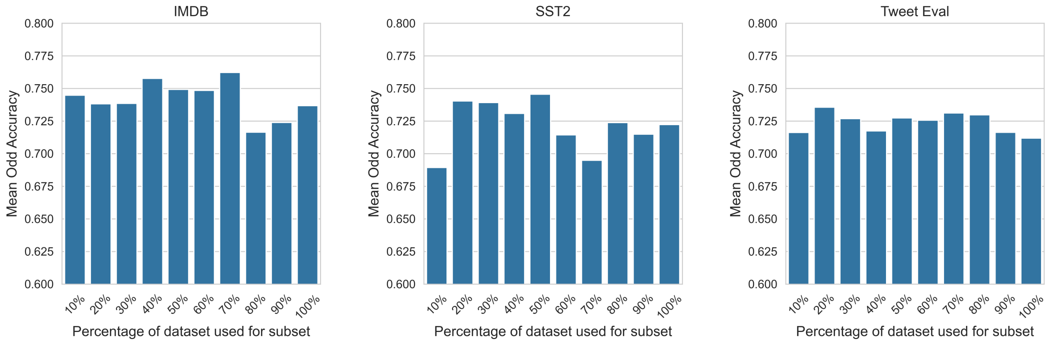

Figure 2: Mean Out-of-distribution accuracy for the subsets of models trained on Sentiment Classification datasets.

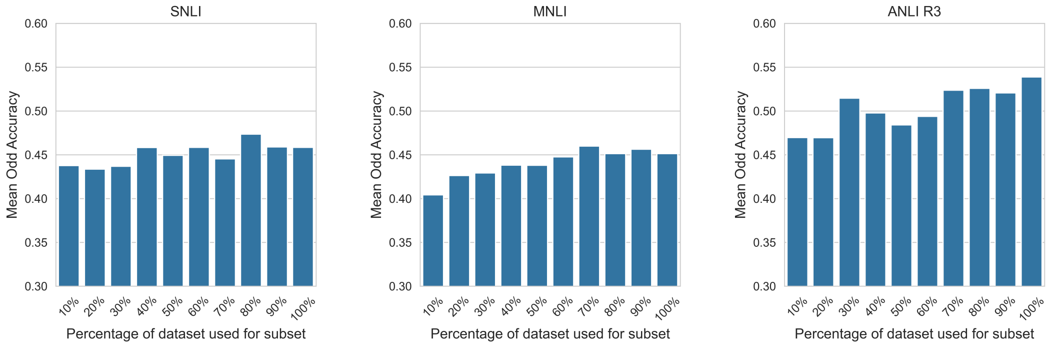

Figure 3: Mean Out-of-distribution accuracy for the subsets of models trained on Natural Language Inference datasets.

a marginal or no increment in the explained variance when selecting a larger proportion of the dataset. While some dataset plots saturate early, for others the plateau come later. This indicates that a smaller subset can explain close to the same variability as explained by a larger subset of the dataset. For sentiment classification, we observe that SST2 has the least variability with the trend plateauing very early with marginal increase in explained variance. For IMDB and TweetEval, we observe similar trends. For natural language inference, we observed that SNLI has the least variability according to our measure, while MNLI shows marginally higher increase. The maximum increase with a later plateau is observed in the ANLI R3 curve. We further study in-distribution and out-of-distribution performance of the models trained on the dataset subsets and its relationship with the magnitude of variability observed for the datasets.

For evaluating model robustness, we analyze out-of-distribution performance of the subset models on different datasets of the same task and plot the mean accuracy. From Figure 2 and Figure 3, we observe that 5 out of 6 models across the 2 tasks that were trained on 100% of the datasets suffer in their generalization capabilities when compared to models trained on a subset of the dataset, having a lower mean out-of-distribution accuracy than one of the smaller subset models. Furthermore, correlating subset model performance with explained variance curves in sentiment classification reveals that the model trained on a subset just before the plateau region demonstrates the highest mean accuracy or closely aligns with the best mean accuracy across all subsets. In the plateau region, performance degradation is observed across datasets. For natural language inference, SNLI and MNLI exhibit similar trends with the explained variance curve guiding optimal subset size, but mean accuracy only minimally degrades in the plateau region. However, ANLI R3 presents a unique scenario, showing a higher accuracy increase with increased dataset variability in the variance curve. Notably, training the model on the complete dataset yields the best performance. We attribute this to the adversarial nature of ANLI R3, where the model learns with more samples on a difficult dataset

Sentiment Classification

	In-Distribution			Out-of-Distribution		
Dataset Name	**10%**	**50%**	**100%**	**10%**	**50%**	**100%**
IMDB	0.849	0.873	0.872	0.745	0.749	0.737
SST2	0.895	0.927	0.952	0.689	0.746	0.722
Tweet_eval	0.749	0.775	0.764	0.716	0.727	0.712

Natural Language Inference

	In-Distribution			Out-of-Distribution		
Dataset Name	**10%**	**50%**	**100%**	**10%**	**50%**	**100%**
ANLI	0.417	0.461	0.451	0.470	0.484	0.539
MultiNLI	0.787	0.825	0.836	0.404	0.438	0.451
SNLI	0.862	0.886	0.893	0.438	0.450	0.458

Table 3: In-distribution and Out-of-distribution results for subset sizes - 10%, 50% and 100%

without overfitting to the distribution. We hope to evaluate better methods of feature construction for tasks with multiple inputs to compute the explained variance curve.

Moreover, our explained variance curve, when considered independently of other datasets, serves as a robust indicator for determining the optimal subset size for random sampling in training a robust model. When compared with other datasets of the same task, it offers valuable insights into dataset quality by assessing variety.

6 CONCLUSION

In this work, we introduce a dataset variability measure - explained variance curve, which provides an indicator on the variability present in a dataset. We empirically show that models trained on full dataset suffer in their generalization power compared to model trained on its subsets and explained variance curve provides a method to determine the right subset size without training for a sampling technique that could help train more robust models. Training on a subset would further reduce training time and hence, resoure utilization.

7 FUTURE WORK

Our study lays a solid foundation for further exploration in two distinct but complementary directions: expanding the breadth of our analysis across more diverse models and datasets, and deepening our investigation into different domains beyond NLP, such as Computer Vision. Another potential future work direction is to evaluate the efficacy of explained variance curve on pre-training data of Language Models where it would assist in selecting a representative subset with better generalization capabilites as show in Abbas et al. (2023).

REFERENCES

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication, 2023.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information, 2022.

Charan Gowda, Anirudh, Akshay Pai, and Kumar A Chaithanya. Twitter and reddit sentimental analysis dataset, 2019.

Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16. International World Wide Web Conferences Steering Committee, April 2016. doi: 10.1145/2872427.2883037. URL `http://dx.doi.org/10.1145/2872427.2883037`.

Eeshaan Jain, Tushar Nandy, Gaurav Aggarwal, Ashish V. Tendulkar, Rishabh K Iyer, and Abir De. Efficient data subset selection to generalize training across models: Transductive and inductive networks, 2023. URL `https://openreview.net/forum?id=GKpwIa9wgwR`.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training, 2021.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1015`.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. Good debt or bad debt: Detecting semantic orientations in economic texts, 2013.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In Preslav Nakov and Torsten Zesch (eds.), *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 1–8, Dublin, Ireland, August 2014a. Association for Computational Linguistics. doi: 10.3115/v1/S14-2001. URL `https://aclanthology.org/S14-2001`.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 216–223, Reykjavik, Iceland, May 2014b. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf`.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John F. J. Mellor, Irina Higgins, Antonia Creswell, Nathan McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, L. Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, N. K. Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Tobias Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew G. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem W. Ayoub, Jeff Stanway, L. L.

Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446, 2021. URL https://api.semanticscholar.org/CorpusID:245353475.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pp. 502–518, 2017.

Rahul Singh, Karan Jindal, Yufei Yu, Hanyu Yang, Tarun Joshi, Matthew A. Campbell, and Wayne B. Shoumaker. Robustness tests of nlp machine learning models: Search and semantically replace, 2021.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1170.

Jun Suzuki, Heiga Zen, and Hideto Kazawa. Extracting representative subset from extensive text data for training pre-trained language models. *Inf. Process. Manage.*, 60(3), may 2023. ISSN 0306-4573. doi: 10.1016/j.ipm.2022.103249. URL https://doi.org/10.1016/j.ipm.2022.103249.

Aarne Talman and Stergios Chatzikyriakidis. Testing the generalization power of neural network models across nli benchmarks. In *BlackboxNLP@ACL*, 2018. URL https://api.semanticscholar.org/CorpusID:54062472.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101.

Jin Zhang, Guannan Liu, and Ming Ren. Finding a representative subset from large-scale documents. *J. Informetrics*, 10:762–775, 2016. URL https://api.semanticscholar.org/CorpusID:41082048.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. *arXiv:1509.01626 [cs]*, September 2015.